Does not say final report 12/31/73

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION
WASHINGTON, D.C. 20546

2/23/73

REPLY TO ATTN OF: DHC-6/KM:eh

{56 Reports by author, Walsh and {37 reports in RECON under Grant

TO : REE/Mr. Charles Pontious

FROM : DHC-6/Contract Administration Branch
Headquarters Contracts Division

SUBJECT: NASA Grant NGR 44-007-028 12/31/69

Records of this branch indicate that a final report has not been received.

DRA

Since this grant is physically completed, please indicate by endorsement below that the work has been satisfactorily completed and that the final report requirement has been satisfied.

Joseph T. Davis

I certify that work under subject grant has been satisfactorily completed.

C. E. Pontious
Signature

2/26/73
Date

N73-23660

(NASA-CR-131989) EXACT INTERVALS AND TESTS FOR MEDIAN WHEN ONE SAMPLE VALUE POSSIBLY AN OUTLINER Final Report (Southern Methodist Univ.) 7 p HC $3.00 CSCL 12A G3/19

Unclas 17613

I

SOUTHERN METHODIST UNIVERSITY

RESEARCH ADMINISTRATION
DALLAS, TEXAS 75222
TELEPHONE 214-692-2031

February 14, 1973

Mr. Joseph T. Davis, Grants Officer
National Aeronautics and Space Administration
Washington, D. C. 20546

Attn: DHC-2(EHawley)

Subject: NASA Grant NGR 44-007-028
SMU 83-21

Per your request enclosed is copy of the paper which Dr. John E. Walsh, Principal Investigator under subject grant, presented in Dublin, Ireland, at the time of his death. Copy of this paper is also being forwarded to Mr. Charles E. Pontius, Code REE, with copy of this letter.

If this will satisfy your requirement for a final technical report, our Accounting Department will now prepare a final fiscal report preparatory to termination of this grant.

Truman F. Cook
Director

cc: Mr. Charles E. Pontius
Code REE

II

FINAL REPORT

EXACT INTERVALS AND TESTS FOR MEDIAN WHEN ONE "SAMPLE" VALUE POSSIBLY AN OUTLIER

Grace J. Kelleher
University of Texas at Arlington*

John E. Walsh
Southern Methodist University**

ABSTRACT

Available are n independent observations (continuous data) that are believed to be a random sample. Desired are distribution-free confidence intervals and significance tests for the population median. However, there is the possibility that either the smallest or the largest observation is an outlier. Then, use of a procedure for rejection of an outlying observation might seem appropriate. Such a procedure would consider that two alternative situations are possible and would select one of them. Either (1) the n observations are truly a random sample, or (2) an outlier exists and its removal leaves a random sample of size n-1. For either situation, confidence intervals and tests are desired for the median of the population yielding the random sample. Unfortunately, satisfactory rejection procedures of a distribution-free nature do not seem to be available. Moreover, all rejection procedures impose undesirable conditional effects on the observations, and also, can select the wrong one of the two above situations. Such difficulties could be bypassed if intervals and tests are used that simultaneously apply to both situations, i.e. if a confidence coefficient, or significance level, has the same value for both situations. It is found that two-sided intervals and tests based on two symmetrically located order statistics (not the largest and smallest) of the n observations have this property.

*Also allifiated with Computer Aid Companies, Inc.
**Research partially supported by NASA Grant NGR 44-007-028, Department of Labor Grant 31-46-70-07, and Mobil Research and Development Corporation. Also associated with ONR Contract N00014-68-A-0515.

INTRODUCTION AND DISCUSSION

The data are n independent observations that are continuous data and are believed to be a random sample. The order statistics of these observations are

$$x(1) \leq x(2) \leq \ldots \leq x(n).$$

Distribution-free confidence intervals and significance tests are desired for the median θ (not necessarily unique) of the population sampled. However, the possibility exists that x(n) is an outlier, or the possibility exists that x(1) is an outlier. That is, x(n) is so much larger than the other observations that there is doubt that it was produced by the population that produced the other n-1 observations. Alternatively, x(1) is so much smaller than the other observations that there is doubt that it came from the population that yielded the other n-1 observations.

When such a doubt exists, use of a procedure for deciding on the rejection of an outlying observation might seem appropriate. A standard rejection procedure would consider that two situations are possible and, on the basis of the observations, would select one of these two situations (as that which occurs).

The n observations are truly a random sample for one of the two situations (with the median θ of the associated population being investigated). The doubtful observation is an outlier for the other situation. More specifically, the population yielding the suspected outlier is different from the population yielding the other n-1 observations, and in such a way that removal of the outlier leaves a random sample of size n-1. In addition, the population for the random sample obtained under these conditional circumstances is considered to be the same as the population that unconditionally yielded these n-1 observations. Then,

distribution-free intervals and tests are desired for the median θ of the population yielding the sample of size n-1 (for the situation where the doubtful observation is an outlier). Also, when x(n) is an outlier, x(1),..., x(n-1) are the order statistics of the sample of size n-1, while x(2),..., x(n) are the order statistics of this sample when x(1) is an outlier.

Unfortunately, development of a satisfactory procedure for rejection of an outlier is a formidable problem for distribution-free cases. What represents a substantial deviation from the other observations depends strongly on the distribution tail (which can be of any continuous form in the distribution-free cases). Even if a satisfactory rejection procedure could be developed, its use would involve important difficulties. First, the wrong one of the two situations might be selected. Second, use of the rejection procedure would introduce undesirable conditional effects on the probability properties of the observations. For example, suppose that the n observations are truly a random sample. They will no longer be a random sample after being subjected to the rejection procedure, even if the correct situation is selected. That is, only those sets whose n observed values satisfy one or more requirements imposed by the procedure are considered to be random samples.

A more attractive approach would be to use intervals and tests that apply simultaneously to both situations. That is, a confidence interval has the same confidence coefficient for the two situations. Also, a test has the same significance level for both situations. Fortunately, intervals and tests with this property can be developed. In fact, the well-known equal-tail sign tests, and the corresponding two-sided confidence intervals, are shown to have this property (when x(1) and x(n) are

not used). This is the case whether x(n) could be an outlier or whether x(1) can be an outlier. For convenience of presentation, only the confidence intervals are explicitly considered. However, the property for the corresponding test follows in a direct fashion, since the tests can be obtained directly from the intervals.

If the n observations were truly a random sample, the well-known confidence intervals defined by

$$P[x(i) \leq \theta \leq x(n+1-i)] = 1 - (\tfrac{1}{2})^{n-1} \sum_{j=0}^{i-1} \binom{n}{j} \qquad (1)$$

are applicable. These are the confidence intervals considered (for $2 \leq i < n/2$). The relationship (1) is found to hold when x(1) is an outlier and also, when x(n) is an outlier. Verification of this property is given in the next section.

VERIFICATION

Only the situation where x(1) is an outlier receives consideration. A similar method provides verification that (1) holds when x(n) is an outlier.

In general, the value of $P[x(i) \leq \theta \leq x(n+1-i)]$ can be expressed as unity minus

$$P[x(i) > \theta] + P[x(n+1-i) < \theta].$$

When x(1) is an outlier, x(2) becomes the smallest observation, etc. and

$$P[x(i) > \theta] = (\tfrac{1}{2})^{n-1} \sum_{j=0}^{i-2} \binom{n-1}{j},$$

$$P[x(n+1-i) < \theta] = (\tfrac{1}{2})^{n-1} \sum_{j=0}^{i-1} \binom{n-1}{j},$$

with their sum being

$$(\tfrac{1}{2})^{n-1}\sum_{j=0}^{i-1}\left[\binom{n-1}{j} + \binom{n-1}{j-1}\right],$$

where $\binom{n-1}{-1}$ is zero. However, $\binom{n-1}{0} = \binom{n}{0}$ and

$$\binom{n-1}{j} + \binom{n-1}{j-1} = \binom{n}{j}$$

for $1 \le j < i$. Thus, the value of $P[x(i) \le \theta \le x(n + 1 - i)]$ is

$$1 - (\tfrac{1}{2})^{n-1}\sum_{j=0}^{i-1}\binom{n}{j},$$

which is the value of (1).

It is to be noticed that $P[x(i) > \theta]$ does not differ much from $P[x(n + 1 - i) < \theta]$ when i is of at least moderate size (ordinarily implies that n is at least moderately large. A desirable feature of the results presented is that the probability can be accurately determined for each tail of an interval or test.